MTA SZTAKI, Országos Közegészségügyi Intézet,
MTA Matematikai Kutató Intézet

Többszörös veleszületett rendellenességek statisztikai vizsgálata

Bolla Mariann, Czeizel Endre, Telegdi László,
Tusnády Gábor

Bevezetés

Hazánkban 1970-ben szervezték át a veleszületett rendellenességek /VR/ nyilvántartását. Az adatok szerint az 1971-75 között született 844 886 ujszülöttnek kb. 3,5 %-ánál észleltek olyan rendellenességet, ami már a terhesség idején kialakul. A leggyakoribb ilyen VR-ek: a veleszületett csipőficam /Cd/, szív- és keringési rendellenességek /C/, sérvek /He/, agynélküliség nyitott gerinccel /As/ és a dongaláb /T/. /A nyilvántartott 40 rendellenesség - néhány esetben rendellenesség-csoport - mindegyikét egy kóddal láttuk el./

Megfigyelhető, hogy bizonyos VR-ek gyakran fordulnak elő másokkal. Vizsgálódásaink középpontjába épp az ilyen többszörös esetek kerültek. Itt szeretném megemliteni, hogy a már ismert szindrómákat /pl. Down-, Edwards-,

Patau-szindrómák/ kizártuk az alapadatok közül.

Véletlen vagy szisztematikus társulás?

Miért van szükség a VR-ek együttes vizsgálatára? Röviden talán választ ad már a következő arány is: azok közül az ujszülöttek közül, akiken megfigyeltek valamely rendellenességet, 13 %-nál legalább még egyet is észleltek. Ez a halmozódás is mutatja, hogy a VR-ek gyakran társulnak. Ha sikerülne köztük kapcsolódási csoportokat felfedezni, az nemcsak nyilvántartásukat tenné könnyebbé, hanem esetleg a VR-ek kialakulására, genetikai hátterére, illetve az anyát ért káros környezeti hatásokra vonatkozóan is mondana valamit.

Az első kérdés, ami ezzel kapcsolatban felvetődhet: véletlen vagy szisztematikus társulásokról van-e itt szó?

A függetlenség hipotézisének elvetése

Jelölje $p(k)$ a k-adik rendellenesség előfordulásának valószinüségét, $P(G)$ pedig a $G = k_1, k_2, \ldots, k_s$ rendellenesség-kombináció valószinüségét. Ha a VR-ek függetlenek lennének, akkor

$$P(G) = Q \prod_{k_i \in G} q(k_i) \qquad (1)$$

lenne, ahol $q(k) = p(k)/(1 - p(k))$ és $Q \prod_{k=1}^{40} (1 - p(k))$.

Jelölje $\mathcal{P}(s)$ az összes s elemszámu csoport valószinüségének összegét:

$$\mathcal{P}(s) = \sum_{|G|=s} P(G) . \qquad (2)$$

Ha q(k)-k kicsik /mint esetünkben is/, akkor $\mathcal{P}(s)$ közelitőleg Poisson-eloszlást követ

$$\mathcal{P}(s) = Q \sum_{|G|=s} \prod_{k_i \in G} q(k_i) \sim \frac{Q_1}{s!} \left(\sum_{k=1}^{40} q(k) \right)^s , \qquad (3)$$

ahol $Q_1 = \exp\left(- \sum_{k=1}^{40} q(k)\right)$.

Azonban a $\mathcal{P}(s)$-eknek megfelelő gyakoriságértékeket s = ,2,...,7-re kiszámolva azt tapasztaltuk, hogy ez nem teljesül.

Statisztikai vizsgálatok az adatrendszeren

Statisztikai vizsgálódásainkat két lépésben végeztük:

1. Olyan matematikai modellt kerestünk, amelyben az egyes VR-ek valószinüsége és együttes előfordulásaik valószinüsége is meghatározható.

2. Az igy kapott korrelációs strukturát analizáltuk, a VR-ek között kapcsolódási csoportokat kerestünk.

Célunknak megfelelően az adatrendszer azoknak az újszülötteknek a számát tartalmazza, akiken egy-egy rendellenesség-kombináció - és csak ez a rendellenesség-kombináció - megfigyelhető. A lehetséges 2^{40} kombináció közül természetesen csak a nem - 0 gyakoriságuakat töltöttük be a gépbe, de még igy is elég bonyolult volt az összes adatot ugy elhelyezni a memóriában, hogy a számokat gyorsan ki lehessen keresni. Ezt külön szubrutinok végzik, amit itt nem részletezek.

A GAMT-modell

Kiindulásul a Gaussian Additive Multifactorial Threshold /GAMT/ modellt választottuk /2/. Eszerint minden egyes VR mögé egy normális eloszlásu háttérváltozót, un. hajlamot képzelünk, és egy VR pontosan akkor következik be, ha ez a hajlam meghalad egy adott - a populációra jellemző - értéket, a küszöböt. Ezt is feltettük, hogy a rendellenességek hajlamai együttesen többdimenziós normális eloszlást követnek. A hajlamokat L_1, L_2, -, L_{40}-el, a küszöböket pedig T_1, T_2, ..., T_{40}-el jelölve a G rendellenesség-kombináció bekövetkezésének valószinüsége:

$$P(G) = P(L_k \geq T_k, \quad k \in G) . \qquad (4)$$

A küszöbök az egyszeres előfordulások gyakoriságaiból becsülhetők. A várható értékeket 0-nak véve /az egy li-

neáris transzformációval elérhető/ a hajlamok együttes eloszlását korrelációs mátrixuk már egyértelmüen meghatározza. A $\varrho_{ij} = E(L_i L_j)$ korrelációkat a páros előfordulásokból becsültük.

Egy VR bekövetkezését lehetne a megfelelő hajlamnak egy - a hajlamok 40 dimenziós terében adott - halmazba való esésével reprezentálni:

$$P(G) = \mathcal{P}(L_k \in S_k, \quad k \in G) \; . \qquad (5)$$

Igy a modell általánositásához jutunk.

A modell tesztelése

A GAMT-modell illeszkedését az fejezi ki, hogy a legalább háromszoros előfordulásokat mennyire adja viszsza a kapott korrelációs struktura. Az ilyen kombinációk valószinüsége elvileg könnyen számolható. A gyakorlatban azonban az alkalmazott módszer nagymértékben függ a dimenziós-számtól és a korrelációs mátrixnak, illetve részmátrixainak tulajdonságaitól.

A 4-7-szeres esetek tesztelése Monte Carlo módszert alkalmaztunk. Generáltunk 40 dimenziós normális eloszlásu valószinüségi változókat az adott korrelációs mátrixszal. Egy ilyen egy született gyereknek felelt meg, melynek komponensei az ujszülöttnek az egyes rendellenességekre való hajlamai. Minden egyes esetben megnéztük, hogy

mely komponensek vannak a megfelelő küszöb felett, vagyis, hogy az ujszülöttnek milyen VR-ei vannak. A program azonban nagyon lassu volt a kis valószinüségek miatt, ezért nem sikerült kellő nagy mintát generálni.

3-dimenzióban a Pearson-féle sorfejtést használtuk, melynek általános alakja s dimenzióban /1/:

$$P(L_1 \geq T_1,\ L_2 \geq T_2, \ldots, L_s \geq T_s) = \prod_{\nu=1}^{s} Q(T_\nu) + \prod_{\nu=1}^{s} \varphi(T_\nu) \cdot$$

$$\sum_{q=0}^{\infty} \sum_{\substack{\alpha_{ij} \geq 0 \\ \alpha_{ij} \text{ egész} \\ \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \alpha_{ij} = q}} \left(\prod_{i=1}^{s-1} \prod_{j=i+1}^{s} \frac{\rho_{ij}^{\alpha_{ij}}}{\alpha_{ij}!} \right) \prod_{\nu=1}^{s} H_{\beta_\nu - 1}(T_\nu), \tag{6}$$

ahol $\beta_\nu = \sum_{i=1}^{\nu-1} \alpha_{i\nu} + \sum_{j=\nu+1}^{s} \alpha_{\nu j}$, $R = (\rho_{ij})_{i=1,j=1}^{s\ \ s}$ a korrelációs mátrix,

$$\varphi(x) = (1/\sqrt{2\pi})e^{-x^2/2}, \quad Q(x) = \int_x^{+\infty} \varphi(t)dt, \quad H_k$$

pedig a k-adik Hermite-polinom. A sor abszolut- és egyenletesen konvergens egész R^s-en, ha R-I sajátértékei abszolut értékben 1-nél kisebbek, I jelöli az s x s -es identitás mátrixot. Ennek elégséges feltétele, hogy a korrelációk abszolut értékben 1/s-1 -nél kisebbek legye-

nek. Ez 3 dimenzióban teljesül nálunk, magasabb dimenziókban azonban egyre több az ennél nagyobb korreláció.

A fenti képletet a sor átrendezésével és konvergenciát gyorsitó konstansok bevezetésével programoztuk be:

$$P(L_1 \geq T_1,\ L_2 \geq T_2, \ldots, L_s \geq T_s) = \prod_{\nu=1}^{s} Q(T_\nu) + \prod_{\nu=1}^{s} \varphi(T_\nu) \cdot \tag{7}$$

$$\sum_{\alpha_{12}=0}^{\infty} \sum_{\alpha_{13}=0}^{\infty} \cdots \sum_{\alpha_{s-1,s}=0}^{\infty} \prod_{\nu=1}^{s} \beta_\nu \Big(\prod_{i=1}^{s-1} \prod_{j=i+1}^{s} \frac{\varrho_{ij}^{\alpha_{ij}}}{\alpha_{ij}!} \Big) \prod_{\nu=1}^{s} \frac{H_{\beta_\nu - 1}(T_\nu)}{\beta_\nu !} .$$

A valószinüségeket a program kb. 0,1 sec alatt számolja 0.000001 nagyságrendü hibával.

Jelölje $E_T(G)$ a G rendellenesség-kombináció várt értékét és legyen $E(k) = \sum_{G: k \in G, |G|=3} E_T(G)$, $O(k)$ pedig a megfelelő megfigyelt érték. A kapott és a megfigyelt értékek eltérése egyes rendellenességekre nézve nem szignifikáns, viszont a várt értékek csaknem mindig nagyobbak a tapasztaltaknál /ld. 1. táblázat/. Egyes csoportokra nézve ez már jelentős eltérést adhat, ami egyben lehetőség is szindrómák keresésére.

Nézzük meg, mi lehet az eltérés oka?

a./ Lehet egyszerü számolási hiba az adott gépidő és hibakorlát mellett.

b./ Lehet a modellünktől való eltérés is, ami most azt jelenti, hogy a kettőnél többszörös esetekre már nem illeszthető a többdimenziós normális eloszlás. Ennek hátterében az állhat, hogy a normalitás feltevése a multifaktoriális kóreredetnek felel meg: eszerint a VR-ek sok kis gén hatására, továbbá környezeti hatásokra alakulnak ki, és mint tudjuk, sok független, azonos eloszlású valószinüségi változó eredője normális eloszlású lesz. De a kromoszomák gén-részletei végesek, sőt az is lehet, hogy egy rendellenességet meghatározó gének száma elenyészően kicsi, esetleg egyetlen "hibás" gén is okozhat VR-t. A várt és megfigyelt gyakoriságok közti szisztematikus eltérések is arra utalnak, hogy monogén hatások is vannak jelen.

c./ Talán nem világos, hogy most poligén- vagy monogén hatások dominálnak-e, de ez még orvosi körökben is vitatott kérdés. Nem is lényeges, ha a következő feltevéssel élünk: Minden egyes VR-nek van egy multifaktoriális és egy unifaktoriális változata. Azaz van egy változata, ami a többi rendellenesség hasonló változataival együtt poligén módon öröklődik, ugyanakkor a másik változata - mint valamely szindróma része - egy vagy néhány "hibás" gén hatására ugyanilyen formában képes manifesztálódni. Igy éppen azok a rendellenesség-kombinációk, melyek gyakorisága szignifikánsan eltér a GAMT-modell

alapján várt értékektől, alkothatnak szindrómákat.

d./ Hipotézisünk tisztázását segitené, ha ismernénk a VR-eknek a méhen belüli fázisban való kialakulását. Lehet ugyanis, hogy bizonyos rendellenességek megléte már automatikusan maga után vonja a terhesség további szakaszában egyéb rendellenességek megjelenését, és ez is okozhatja, hogy a megfigyelt értékek meghaladják a vártakat.

A korrelációs struktura analizise

Eddigi számolásaink a modellből kapott korrelációkon történtek. A korrelációk alapján hajtottunk végre cluster analizist is, ami a távolságok választásának megfelelően különböző csoportokat adott. A kovarianciastruktura analizisének egzaktabb módszere a faktoranalizis. /A kiszöbök, a sajátértékek és az első 8 faktor a második táblázatban láthatók./ A számitógépes program a főfaktormódszeren alapult.

Az első sajátérték elég nagy ahhoz, hogy egy 1-faktoros modellt vegyünk fel. Ez azért is előnyös, mert ekkor a P(G) valószinüségek egyszerüen számolhatók:

$$P(G) = \int_{-\infty}^{+} \varphi(t) \prod_{k \in G} Q(v_k)\, dt ,$$

ahol

$$v_k = T_k - t\lambda_k/(1-\lambda_k^2)^{1/2}, \quad \lambda_1,\lambda_2,\ldots,\lambda_{40} \qquad (8)$$

pedig az első faktor sulyai.

1. Táblázat: A rendellenességek előfordulása a 3-as csoportokban

Sor-szám	Kód	Előfordulási gyakoriság		X^2
		Várt	Megfigyelt	
1	C	533.22	586	5.22
2	T	223.16	319	41.16
3	A	222.03	237	1.01
4	Cl	217.15	252	5.59
5	P	271.33	292	1.57
6	As	226.67	234	0.24
7	G	82.08	89	0.58
8	S	50.25	48	0.10
9	Hc	243.04	242	0.01
10	U	93.37	136	19.46
11	Cd	51.65	100	45.26
12	Di	74.06	95	5.92
13	Sh	101.01	85	2.54
14	He	74.77	81	0.52
15	N	18.93	24	1.36
16	Ut	164.66	185	2.51
17	R	126.47	132	0.24
18	Ea	230.84	235	0.08
19	F	147.92	171	3.60
20	Cp	114.17	123	0.68
21	O	110.73	113	0.05
22	Rs	70.30	73	0,10
23	Mc	135.17	116	2.72
24	Ex	121.61	112	0.76
25	Ey	77.49	89	1.71
26	Ds	53.98	68	3.64
27	M	143.95	120	3.98
28	Ps	2.80	5	1.73
29	Ra	75.91	64	1.87
30	Ph	102.16	80	4.81
31	K	83.05	68	2.73
32	Is	13.40	10	0.86
33	Ig	12.08	20	5.19
34	L	6.83	7	0.00
35	Br	5.69	4	0.50
36	Tu	2.29	2	0.04
37	Mr	4.04	5	0.23
38	Eo	13.85	6	4.44
39	Ns	3.87	10	9.71
40	Au	0.10	-	0.10

2. Táblázat: A korrelációs mátrix faktoranalizise

kód		küszöb	saját érték	faktorok							
				1	2	3	4	5	6	7	8
1	C	2.59	10.18	.70	.07	.14	.08	-.03	.09	-.06	-.06
2	T	2.86	2.79	.54	-.10	-.13	.00	.26	-.09	-.09	-.13
3	A	3.38	2.10	.73	.10	-.23	.25	.26	.13	.01	-.17
4	Cl	3.00	1.99	.54	-.17	-.31	.03	-.08	-.10	-.14	.19
5	P	3.05	1.87	.64	-.18	.03	.00	.29	.32	-.06	-.13
6	As	2.77	1.75	.51	.23	-.19	.09	.12	.41	.18	-.04
7	G	3.81	1.70	.64	.13	-.48	-.17	.12	.32	.13	.13
8	S	3.93	1.69	.67	-.12	-.23	-.38	.12	.01	.26	-.20
9	Hc	3.10	1.44	.64	.07	-.04	.08	-.35	.05	.06	.15
10	U	2.98	1.41	.41	.34	-.17	.18	-.15	.18	.00	-.30
11	Cd	2.49	1.34	.13	-.23	-.07	-.10	.23	.07	.25	.06
12	Di	3.42	1.28	.54	.30	.01	.11	-.38	-.66	.17	-.19
13	Sh	2.97	1.20	.41	-.22	-.07	-.03	.04	.30	-.09	.12
14	He	2.74	1.13	.26	-.17	.26	-.27	.14	-.03	-.39	-.17
15	N	2.77	1.09	.10	-.27	.20	-.18	.23	.25	-.05	.22
16	Ut	3.15	1.01	.57	-.22	-.13	-.09	.07	.17	-.06	-.04
17	R	3.31	0.99	.61	-.03	.01	.30	.12	.09	-.03	.40
18	Eo	3.41	0.91	.70	-.42	-.04	.18	-.14	-.04	.13	-.04
19	F	3.44	0.85	.67	-.23	.30	-.18	-.07	-.03	-.30	.12
20	Cp	3.27	0.79	.51	-.18	.49	.17	-.04	-.38	.26	-.26
21	O	3.44	0.71	.57	.47	.12	.37	.16	.11	-.16	.14
22	Rs	3.47	0.68	.54	.38	-.03	-.28	-.12	.21	.23	.00
23	Mc	3.47	0.66	.64	-.47	.10	-.34	.03	.16	-.05	-.05
24	Ex	3.42	0.59	.61	.18	-.09	.12	.11	-.22	-.30	.12
25	Ey	3.46	0,50	.48	-.25	.16	.24	-.16	-.07	.22	.17
26	Ds	3.35	0.48	.45	.48	.14	-.14	-.18	.09	-.31	.03
27	M	3.46	0.40	.70	-.07	.16	-.01	-.25	.06	.18	.22
28	Ps	3.34	0.35	.09	.00	.22	.06	-.04	.12	-.40	-.85
29	Ro	3.63	0.33	.57	.37	.25	-.13	.27	.30	.40	-.04
30	Ph	3.76	0.20	.64	-.37	-.14	.00	-.41	-.36	-.23	-.13
31	K	3.57	0.09	.57	.30	-.25	-.28	-.19	-.22	-.36	.08
32	Is	3.54	0,06	.32	-.12	.19	.35	-.17	.45	-.38	-.21
33	Ig	3.73	0,05	.26	-.40	.22	.03	.05	.03	.01	-.35
34	L	3.70	-0.06	.25	-.17	.17	.20	.32	.21	-.09	-.51
35	Br	3.75	-0.10	.32	.25	.45	-.61	.34	-.13	-.14	-.03
36	Tu	3.66	-0.20	.26	-.13	-.52	-.34	.11	-.18	-.15	-.23
37	Mr	3.99	-0.36	.19	.15	-.12	.47	.62	-.26	-.18	.22
38	Eo	3.59	-0.44	.32	.57	.33	-.14	-.12	-.20	.66	.13
39	Ns	3.74	-0.65	.19	.05	.04	.14	.04	-.20	.27	-.49
40	Au	3.83	-0.81	.06	-.10	.26	.00	.03	-.26	.09	-.03

I r o d a l o m

[1] M. S. Taqqu: "Law of the Iterated Logarithm for Sums of Non-linear functions of Gaussian Variables that Exhibit a Long Range Dependence", Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete /1977/, 203-214.

[2] Tusnády, G., Telegdi L., Czeizel E. /1978/: "ML-fitting of multifactorial thershold models", Alk. Mat. Lapok /sajtó alatt/.